2007－NL－181 (6)
2007／9／25

高次元圧縮空間の対話的手法による次元縮小

山崎 啓介　張 諾　渡辺 俊典　古賀 久志

電気通信大学 大学院 情報システム学研究科
〒182-8585 東京都調布市調布ヶ丘 1-5-1
E-mail: keisuke@sd.is.uec.ac.jp

あらまし 高次元特徴空間を用いるテキスト分類等において不必要な次元軸を排除することは計算量などの面から重要な問題である．この問題を解決するためには不要と考えられる次元を見出し，類性能を保つ範囲でその次元を削除していくことを繰り返せば良い．本稿では，まずテキストをその圧縮率ベクトルに着目して特徴付ける方式を示し，そこでの次元縮小法と次元縮小に必要なパラメータ設定を支援する指標関数を提案する．指標関数を参考にしながら対話的に次元縮小を行うことで，分類精度を保ったまま約 50 %の次元縮小が可能となった．
キーワード テキスト分類，特徴量空間，データ圧縮，次元縮小，対話的手法

An Interactive Technique for Feature Space Dimension Reduction

Keisuke YAMAZAKI, Nuo ZHANG, Toshinori WATANABE, and Hisashi KOGA

Graduate School of Information Systems, University of Electro-Communications
1–5–1,Chouhugaoka, Chofu-shi, Tokyo, 182–8585 Japan
E-mail: keisuke@sd.is.uec.ac.jp

Abstract When text classification is implemented in high-dimension space, removing unnecessary dimensions becomes important to reduce computation cost. This problem can be solved by finding out unnecessary dimensions and removing them , keeping the classification power of the space. In this paper, we express texts by compression ratio vectors. After introducing it, we propose an interactive dimension reduction method with an index function. The index function is used to judge whether reduction should be continued or not. By removing unnecessary dimensions by using the interactive processing , we clould achieve 50 % dimension reduction while keeping the classification accuracy of the space.
Key words text classification, feature space, data compression,dimension reduction, interractive processing

1 はじめに

近年，コンピュータやインターネットの普及により文書データの電子化が急速に進行している．そして，増大する文書情報は人間の処理能力をはるかに超え，情報すべてに注意を払うことはもはや不可能であり，多大な「データ（群）」から情報・知識を見つけようとするデータマイニングの研究がなされている [1]．一般に，データマイニングのような処理では高次元の特徴空間におけるクラス分類問題が発生し，計算量の増加など様々な問題が生じる．このような問題を解決するには，高次元特徴空間内で不要と考えられる特徴軸を発見し，削除することが有効と考えられる．しかし，データの性質が不明で，どの特徴軸を削除できるかを判断できない場合が多いため主成分分析や多次元尺度法，さらに計算量 O(N) の FastMap 法 [3] などのようにデータをひとまずユークリッド空間の特徴ベクトルとして表現し，それらの間の変動量 (散らばりのエネルギー) をなるべく保存するように次元縮小を行う方法が研究されてきた．しかし，この方法は各特徴軸の特性までに踏み込めないため，重要な軸を捨ててしま

うといった問題が残る．このような自動的コンピュータ処理の限界をカバーする方策として，ユーザの目視判断により，特徴ベクトルの各要素の特性を把握しながら空間の次元縮小を行える対話的なシステムが有用であると考える．なお，高次元データのクラスタリングのため対話的手法の提案例はあるが [7]，次元縮小のための対話的手法については筆者の調査した範囲では顕著な方式が見当らない．

本稿では，まずテキストデータの特徴を圧縮率ベクトルという形で表現し，ベクトル要素を折れ線グラフで表示することでテキストデータの特徴の可視化を図る．次にこの折れ線グラフを用いたテキストの分類手法を与える．さらに，クラス分類能力を保ったまま特徴次元数を削減する対話的手法を与える．その際，削減処理を続行すべきか否かを判断するための指標関数を導入する．これにより，ユーザは指標関数と折れ線グラフを参考にしながら対話的にクラス分類結果を保ったまま，低次元圧縮率ベクトル空間を作成できる．この手法で，分類能力を保ったまま約 50 %の次元縮小が可能であることを実験で示す．

2 対話的次元縮小法

提案手法の手順を図 1 に示す．先ず，テキストデータを圧縮率ベクトルで表現する．次に，これをクラスタリングし，結果を折れ線グラフで表示する．クラスタリングでは k-means 法のようにクラス総数を入力することなく，自動的に分類を行う．作成された折れ線グラフから目視で縮小候補次元を見出す．次元縮小後の特徴空間による全ベクトルのクラス分類結果の変化を表す指標関数を参照しながら，分類能力が低下しない範囲で次元縮小を行う．以下に，各ステップの詳細について述べる．

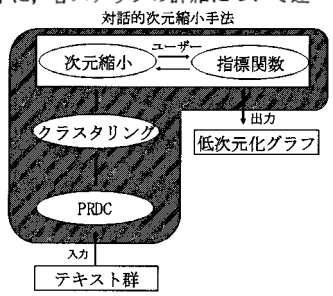

図 1: 本研究法での対話的次元縮小手順

2.1 PRDC(Pattern Representation Scheme Using Data Compression)

提案手法ではデータの特徴を PRDC[4] を用いて表現する．PRDC では，A, B ふたつのデータがあったとき，A を圧縮する際に作られた圧縮辞書を用いて B を圧縮した場合，高圧縮であれば A と B は似ているとみなす．複数個の圧縮辞書を用いて任意のデータを多次元の圧縮率ベクトルとして特徴表現してマルチメディア情報を分析する枠組みが PRDC である．文書データ処理の場合，あらかじめ様々なジャンルの文書群 $T = \{t_1, t_2, ..., t_n\}$ を収集しておき，各文書を圧縮する過程で得られる圧縮辞書群 $D = \{d_{t_1}, d_{t_2}, ..., d_{t_n}\}$ を基底辞書として作成した空間を高次元圧縮率空間とする．入力データ u が与えられた時，基底辞書により圧縮すると，辞書個数分の出力 $U = \{u_{d_{t_1}}, u_{d_{t_2}}, ..., u_{d_{t_n}}\}$ が得られる．圧縮率ベクトルは各出力データ長と元のデータ長を用いて式 (1) のように定義される．高圧縮であるほど次元要素の値は小さくなる．

$$\rho_u = \{\frac{u_{d_{t_1}}}{|u|}, \frac{u_{d_{t_2}}}{|u|}, ..., \frac{u_{d_{t_n}}}{|u|}\} \quad (1)$$

入力データの特徴は圧縮率ベクトルという形で数値化され，分類や類似検索などをこのベクトルを用いて実現できる (図 2). このベクトルをクラスタリングすることで文書データなどの分類を行うことができる [5].

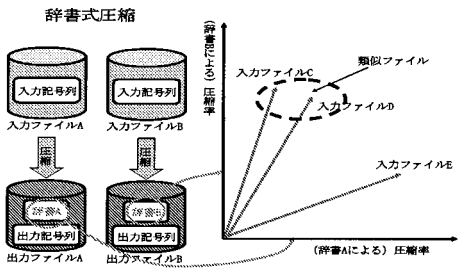

図 2: PRDC による圧縮空間

2.2 データ分類

本研究では，折れ線グラフで文書データの圧縮率ベクトルを表示し (圧縮率グラフ)，文書データと各辞書との関係を参照しながら分類を行う．ここで，k-means 法のようにクラス総数をユーザが指定せず，終了条件となる圧縮率上限値 K(%) のみを与える．提案法は以下のようになる．

データ分類処理

Step.1 全文書から得た基底辞書群により，全文書の圧縮率グラフを作成.

Step.2 全基底辞書について圧縮率が K %以下となっているグラフを探索．(K の初期値は K=0).

Step.3 下記 1, 2 に従ってグラフを分類.

1．発見したグラフが全て新しいならば，それらを

新グループとして登録.

2．発見したグラフ中に一つでも既登録のものがあれば，その所属グループに全ての発見グラフを登録.

Step.4 K の値を $K \leftarrow K + 1\,\%$ として K=75 %まで Step2, Step3 を繰り返す

Step.5 各グラフを所属グループ毎に別の色で表示する.

終了条件 K の値は，日本語テキストの分類を行う際に類似テキストであると判断される上限圧縮率で良いので，文献 [5] の経験値より K=75 %とする.

2.3 対話的次元縮小法と指標関数

PRDC を利用したデータ分類手法を前項で提案したが，より少ない基底辞書でデータ分類が行えれば，この空間を用いた未知データの所在クラス判定等の計算量を小さくできるなどの利点が生まれる．そこで，データ分類処理 (2.2) で得た結果を用いて独立性の高い基底辞書を発見し，空間の次元縮小を行う．但し，次元縮小後の空間では正しく分類できない文書が生じ，分類精度が劣化する可能性がある．そこで，分類誤りとなったデータを「外れ要素」と定義し，その個数 I を表示し，ユーザがこの値を参考にしながら対話的に処理を進めることで，分類精度を低下させずに次元縮小を行うこととする．I を指標関数と呼ぶこととし、これを用いた対話的次元縮小法を以下に示す.

対話的次元縮小処理

Step.1 クラスタリング結果から各クラスタの平均値ベクトル (=平均値折れ線グラフ) を計算し，昇順に基底辞書を整列させる．グラフはクラス毎に色分けしてある.

Step.2 整列された圧縮率グラフを観察し，類似部分から代表的な基底辞書を選出する．選ばなかった基底辞書を捨てることで次元縮小を行う.

Step.3 次元縮小後，残った基底辞書で再度クラスタリングを行う．分類誤り (=前回サイクルまでの分類結果と異なるグラフの個数) を指標関数 I として表示する.

Step.4 更新後の圧縮率空間において次元縮小続行可能性をユーザが判断する (I=0 なら続行可能)，続行可の場合は Step2, Step3 を繰り返す．そうでなければ次元縮小を終了する.

例として 55 個の文書データに対して提案法を適用した結果を図 3 に示す．太矢印で囲まれた数字はその範囲で選択された基底辞書数である．①，③，⑥，⑦ではその範囲に含まれる圧縮率に目視的大差がない．よって，それらから一つの基底のみを選択している．なお，グラフ下部のノコギリ状変動は，自己圧縮率による最小値に由来するものであり，無視する．④，⑤では同じクラスタ内で矢印で示された部分において圧縮率に差が出ているため，それぞれ 2 個，3 個の基底を選択している．②では矢印で示した部分に明確な圧縮率の差があるので，選択基底を 2 個としたが，$I \neq 0$ となり，空間の分類能力が低下することが判った．$I = 0$ とするためには全ての基底を選択しなければならなかったため，②の範囲からは全部の基底 (10 個) を選出した．このように，圧縮率空間の可視化を行うことで，全基底辞書で空間を作成した場合のクラスタリング性能を保持したまま空間を形成する基底 (=次元) を最小化することができる.

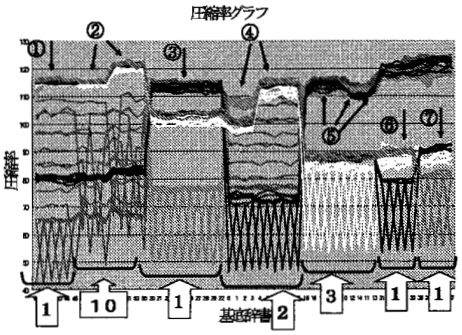

図 3: 基底辞書選出例

3 実験

本節では，モデル文書 (文献 [5] で提案) と実際のニュース記事とに提案法を適用した実験結果を示す．実験 1 ではモデル文書について，実験 2 では実際のニュース記事について実験を行った．ここで，モデル文書とは文書内のフレーズの出現頻度が似ている文書同士が類似文書であるという考えに基づいて，意図的に基本フレーズを指定した回数だけランダムに配置し，それらの間をフレーズに含まれることのない文字 (ノイズ文字) で埋めることで人工的に作成した文書である (図 4).

モデル文書を用いることで，基本フレーズを共有する文書や共有しない文書が圧縮率空間にどのような特徴を示すのかを確認し，実文書に提案手法を適用するための足がかりとした．分類精度の評価には，下記の分類再現率と分類精度を用いた.

$$\text{分類再現率} = \frac{\text{正分類された文書数}}{\text{正解文書数}} \tag{2}$$

$$\text{分類精度} = \frac{\text{正分類された文書数}}{\text{分類対象の文書数}} \tag{3}$$

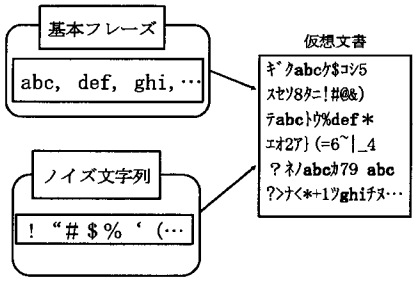

図 4: モデル文書

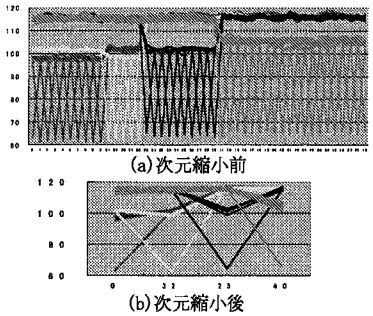

図 5: 次元縮小結果 (横軸は基底辞書，縦軸は圧縮率%)

3.1 実験 1–モデル文書実験

ノイズ 1000 文字に対して基本フレーズを各 20 回ずつ出現させたモデル文書群 (表 1) に対してデータ分類処理 (2.2) の終了条件 K=110 %として提案法を適用し，クラスタリングと次元縮小を試みた．図 5 は，モデル文書群 A，F，I から 10 件ずつと C，D，E から 5 件ずつをとった場合の自動分類，次元縮小結果である．また，得られた次元縮小後の圧縮率空間の分類能力を確認するために，伝統的手法である k-means 法と比較した結果を示した (図 6)．なお，k-means 法での指定クラス総数は，上記 A，F，I，C，D，E 群の基本フレーズの類似性から 3 とした．表 2 に，分類精度を式 (3) とし，提案手法で作成されたクラスタをデンドログラム表示した分類結果と k-means 法の分類結果の比較，及び次元縮小率を示した．

表 1: モデル文書群

モデル文書	基本フレーズ (各 20 回出現)
A1-A10	sport, soccer, baseball
B1-B10	sport, soccer, volleyball
C1-C10	sport, soccer, basketball
D1-D10	animal, dog, cat, pig
E1-E10	animal, dog, cat, bird
F1-F10	animal, dog, cat
G1-G10	weather, snow, sunny
H1-H10	weather, cloudy, sunny
I1-I10	weather, rain, sunny
J1-J10	mount, river, sea

図 5 から次元数を 45 個から 4 個まで縮小することができていることが確認できたので，このデータ群の分類に必要な圧縮率空間の次元数は 4 程度であったことが分かる．また，図 6 では提案手法と k-means 法とで分類結果が異なり，k-means 法では基本フレーズが大略似ている {A，C} と {D，E，F} が各々，同クラスタに分類されたのに対し，提案手法では A，C については基本フレーズが完全に同じ文書同士のクラスタ {A}，{C} が形成された．念のために，提案手法で生成された各クラスタの重心をデンドログラム表示してみると (図 7)，提案法のクラスタが k-means 法でのものと同様の構造をしていることがわかる．よって，分類能力を劣化させず次元縮小を行っていると言える．また表 2 では，k-means 法に対して

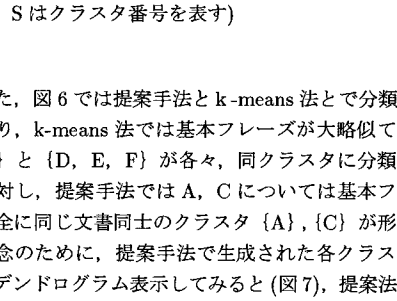

図 6: 提案手法と k-means 法との分類結果 (各点は文書データ，S はクラスタ番号を表す)

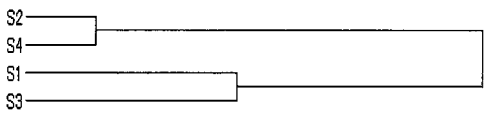

図 7: 図 6 の提案手法のクラスタの重心でのデンドログラム (S はクラスタ番号)

表 2: 平均分類精度と次元縮小率

実験文書数 (クラス総数)	45(3)	100(5)	200(10)
k-means 法	100 %	100 %	100 %
提案手法 (次元縮小前)	100 %	100 %	93 %
提案手法 (次元縮小後)	100 %	99 %	94 %
次元縮小率	91 %	94 %	91 %

常に理想的なクラス総数を与えているためその分類精度 100 %なっているが，実験文書数が 100 程度では提案手法と k-means 法の分類能力に差は無く，分類精度を保ったまま 90 %近くまで次元縮小できている．実験文書数を 200 まで増加させても分類精度の劣化は 10 %未満で，次元数を約 90 %縮小できていることなどがわかる．

3.2 実験 2—実文書実験

表 3 に示す 8 トピックの文書を朝日新聞記事データベース「聞蔵」[6] から各 10 件取得し，実験 3.1 と同様に実験データ数を変化させ，データ分類処理 (2.2) の終了条件 K=75 % [5] とした分類と次元縮小を行った．図 8 はトピックの T1〜T50 の文書群 (5 クラス) を自動分類し，次元縮小した場合の結果である．また，提案手法で対話的に次元縮小した圧縮率空間の分類能力を確認するために，k-means 法と比較した結果を示す (図 9)．なお，k-means 法のクラスタ数は文書のクラス数の 5 とした．表 4 では文書数とトピック数が増えた際の次元縮小率の推移を，表 5，6 では 100 件の文書群を k-means 法により分類した結果と，提案手法で次元を縮小した後に自動分類された結果とを示した．

表 3: トピック

Text ID	概要
T1-T10	履修不足問題
T11-T20	松坂、メジャー移籍
T21-T30	日興コーデ会計不正事件
T31-T40	中越沖地震
T41-T50	アジア杯
T51-T70	参院選
T71-T80	ゴルフ・石川選手
T81-T100	年金不払い問題

表 4: 次元縮小率

実験文書数 (クラス総数)	30(3)	50(3)	100(8)
次元縮小率	60 %	62 %	48 %

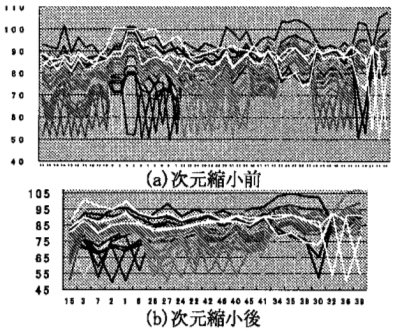

図 8: 次元縮小結果 (横軸は基底辞書，縦軸は圧縮率%)

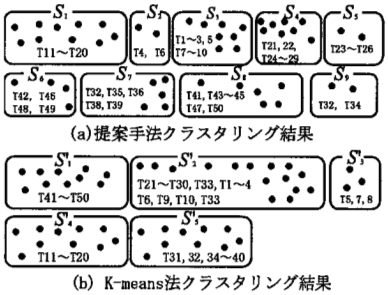

図 9: 提案手法と k-means 法との分類結果 (各点は文書データ，S はクラスタ番号を表す)

図 8 から次元数を 50 個から 19 個にまで削減することができていることが確認できる．よって，このデータ群の分類に必要な圧縮率空間の次元数は 19 程度であったことが分かる．また，図 9 では提案手法と k-means 法とで分類結果が異なり，k-means 法では，S'_1,S'_3,S'_4,S'_5 のように正分類されたクラスタがあるものの，S'_2 では異なるトピックの文書が混ざってしまっている．それに対し，提案手法ではトピックが完全に同じ文書同士のクラスタが形成されたものの，クラス数以上のクラスタが生成されてしまっている．しかし，提案手法で生成された各クラス

図 10: 図 9 の提案手法のクラスタの重心でのデンドログラム (S はクラスタ番号)

タの重心をデンドログラム表示すると (図 10)，提案手法のクラスタがトピックにそった構造をしていることを確認できる．これらの結果より提案手法が，クラスタ数をあらかじめ設定することなく自動分類を行えること，更に分類能力を低化させずに次元縮小を行えることなどを示せた．表 4 を見ると，100 次元の基底を約半分まで縮小できていることが分かり，その分類性能 (表 5，6) を見てみると k-means 法の分類再現率，精度よりも全てのクラスで提案手法が優れており，かつ，表 5 の「*」のクラスのようにどのトピックのクラスであるか判定不可能なものが提案手法ではほとんどなく，トピックに沿った分類ができているといえる．

表 5: k-means 法の分類結果

人手による分類		k-means 法により得られたクラス							
トピック	記事数	クラス 1	クラス 2	クラス 3	クラス 4	クラス 5	クラス 6	クラス 7	クラス 8
履修	10	3					6	1	
松坂	10		10						
日興	10			4				6	
中越	10			1	9				
アジア杯	10					10			
参院選	20			1			16	3	
石川	10					10			
年金	20			1				10	9
分類再現率		30	100	40	90	*	80	*	45
分類精度		100	100	59	100	*	73	*	100

表 6: 次元縮小後の分類結果

人手による分類		次元縮小後の分類クラス							
トピック	記事数	クラス 1	クラス 2	クラス 3	クラス 4	クラス 5	クラス 6	クラス 7	クラス 8
履修	10	10							
松坂	10		10						
日興	10			10					
中越	10				7				
アジア杯	10					9			
参院選	20						15		
石川	10							9	
年金	20						2		17
分類再現率		100	100	100	70	90	75	90	85
分類精度		100	100	100	100	100	88	100	100

4 まとめ

本稿では，テキストマイニング等に有用な PRDC 法での圧縮率空間の分類能力を保存したまま，その次元数を縮小する対話的手法を提案した．提案手法は，適用対象を圧縮率空間に限定したものではあるが，k-means 法と比較して，クラス総数のような，結果として得たい量の事前設定を必要としないこと，対話的次元縮小によって分類能力を保持したまま特徴空間の次元縮小が行えることなどが特徴である．モデル文書と実文書を用いた実験によって，これらの特技を実証した．今後の課題として，大規模データでの検証や，次元縮小時の基底辞書選択の改良による分類精度の向上，などを考えている．

謝辞 本研究は科学研究費基盤研究 (C) 課題番号 19500076 の支援を受けて行った．

参考文献

[1] 石川 慎也，データマイニングはデータで決まる．データをいかに取り，いかに解析するか（データマイニングの宝箱 [http://www5.ocn.ne.jp/shinya91/])，株式会社 産学共同システム研究所.

[2] 神嶌 敏弘，"データマイニング分野のクラスタリング手法 (2) － 大規模データへの挑戦と次元の呪いの克服 －"，人工知能学会誌，vol. 18, No.2, pp. 170-176 , 2003.

[3] Faloutsos, C. and Lin, K.-I.: FastMap: A FastAlgorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets, in Proc. of the ACMSIGMOD Int ' l Conf. on Management of Data, pp. 163-174, 1995.

[4] Toshinori Watanabe, Ken Sugawara, and Hiroshi Sugihara, "A New Pattern Representation Scheme Using Data Compression," IEEE Trans. PAMI, Vol. 24, No. 5, pp. 579-590, May, 2002.

[5] 松崎大輔，渡辺俊典，古賀久志，張 諾，"圧縮性に着目した文書の関係分析手法"，情報処理学会 第 84 回情報学基礎研究会，pp. 51-55, 2006.

[6] http://database.asahi.com/library/.

[7] 中村 朋健，上土井 陽子，若林 真一，吉田 典可，"クラスタリング結果の特徴抽出を用いる高次元データの対話的クラスタリング"，情報処理学会論文誌,Vol. 47, No. SIG-19, pp. 28-41, 2006.